



NVIDIA DGX GB200

생성형 AI를 위한 고급 AI 인프라



수조 개 파라미터 시대의 도래

이제는 규모에 상관없이 많은 기업들이 생성형 AI를 활용하여 챗봇 또는 코파일럿 개발, 콘텐츠 개인화, 신약 개발 가속화, 시각적 애플리케이션 제작 등 다양한 작업을 수행하고 있습니다. 수조 개 파라미터의 페타바이트급 데이터로 훈련하는 파운데이션 모델들도 어렵지 않게 찾을 수 있습니다. 이러한 고성능의 차세대 AI 모델들로 새로운 아이디어를 더욱 효율적으로 반복 실행하고, 결과 도출 시간을 단축하며, 실시간에 가깝게 추론하기 위해서는 수천 개의 GPU를 활용한 훈련 및 추론 인프라가 필요합니다.

엔터프라이즈급 생성형 AI 인프라

NVIDIA DGX™ GB200은 기업에서 전례 없는 성능을 확보하고 가동 시간을 예측 가능하게 하여, 자사의 AI 활용률과 생산성을 극적으로 향상시키고, 투자 대비 수익(ROI)을 높일 수 있게 돕습니다. 이는 AI의 성능, 신뢰성, 확장성 측면에서 새로운 기준을 제시합니다.

DGX GB200은 NVIDIA DGX SuperPOD를 사용하여 수만 개의 GPU로 확장 가능하며, 효율적인 수냉식 랙 스케일 설계를 바탕으로 NVIDIA GB200 Grace Blackwell 슈퍼칩을 활용하여 고급 생성형 AI 애플리케이션에서 필요로 하는 최첨단 AI 모델들을 처리합니다. DGX GB200은 초대형 생성형 AI 훈련 및 추론 워크로드에 적합한 극한의 성능과 안정적인 가동 시간을 제공하는 목적 기반의 솔루션입니다. NVIDIA의 자체 내부 클러스터 설계를 기반으로 구축되어, 엔터프라이즈 AI 인프라로는 최초로 풀스택 복원력 기능을 제공하여, 기업이 운영 복잡성이 아닌 혁신에 집중할 수 있도록 합니다.

개발자 생산성의 극대화

DGX GB200은 AI 인프라에 풀스택 복원력을 제공합니다. 지능형 제어 플레인으로 하드웨어, 소프트웨어, 데이터 센터 인프라 전반에 걸친 수천 개의 데이터 포인트를 지속적으로 트래킹하여 운영 연속성과 데이터 무결성을 보장합니다. 예비 하드웨어와 강력한 체크포인트를 사용하여 페일오버(Failover)를 자동화하고 재시작 매커니즘을 활용하여, 시스템 관리자의 부재 시에도 중단 없이 운영이 가능합니다.

주요 특징

- > NVIDIA GB200 Grace™ Blackwell 슈퍼칩 기반
- > NVIDIA DGX SuperPOD를 통해 수만 개의 GB200 슈퍼칩으로 확장 가능
- > 72개의 NVIDIA Blackwell GPU를 NVIDIA® NVLink™로 하나로 연결
- > 효율적인 수냉식 랙 스케일 설계
- > NVIDIA 네트워킹
- > NVIDIA AI Enterprise 및 NVIDIA Mission Control 소프트웨어 활용

생성형 AI를 위한 슈퍼컴퓨팅

DGX GB200은 최첨단 파라미터 모델을 훈련하고 지연시간에 민감한 추론을 수행 가능하도록 수만 개의 NVIDIA Grace Blackwell 슈퍼칩으로 확장 가능합니다.

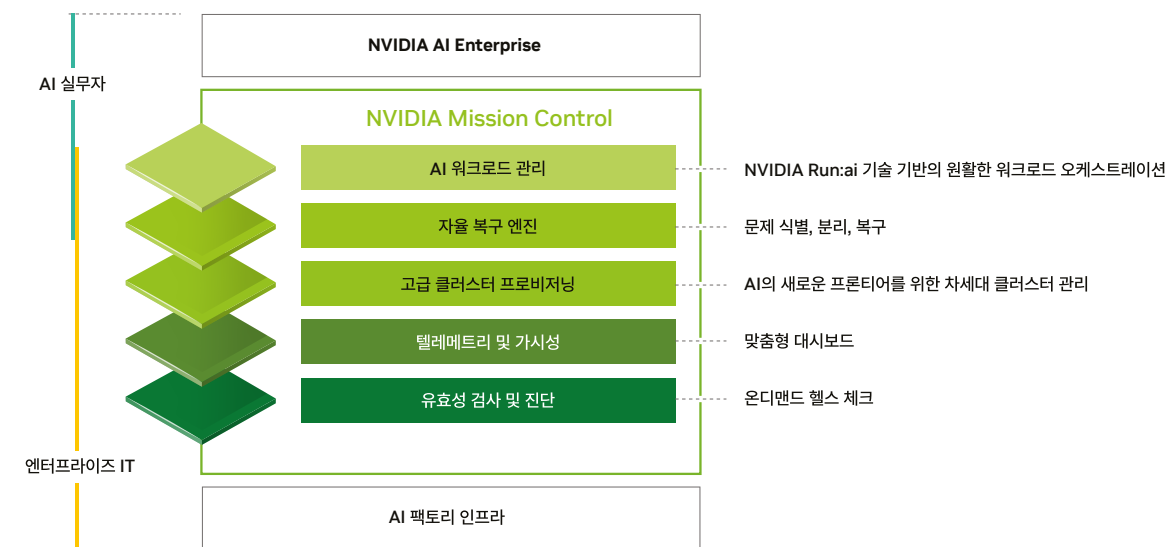
특히 거대 언어 모델(LLM)에 적합하며, 각각의 시스템은 NVIDIA Grace CPU 36개와 NVIDIA Blackwell GPU 72개를 탑재하고, 이를 5세대 NVIDIA NVLink로 하나의 시스템처럼 연결하여, 1.4 exaFLOPS의 AI 성능과 30TB 고속 메모리, 130TB/s의 양방향 GPU 대역폭 제공합니다. 이제 기업에서는 DGX GB200을 사용하여 현재 그리고 미래의 초대형 생성형 AI 모델들을 손쉽게 훈련하고 추론할 수 있습니다.

NVIDIA Grace Blackwell 기반의 구축

DGX GB200을 구동하는 NVIDIA Grace Blackwell 슈퍼칩은 혁신적인 4나노미터(nm) 공정과 5세대 NVLink, 2세대 트랜스포머 엔진을 탑재하고, 수냉식 랙 스케일 설계와 통합되어, 세계에서 가장 효율적인 생성형 AI를 위한 AI 슈퍼컴퓨터를 구축할 수 있습니다. 슈퍼칩 1개는 고성능 NVIDIA Blackwell GPU 2개와 NVIDIA Grace CPU 1개로 구성되어 있으며, 각각의 Blackwell GPU는 GPU 간 연결을 위해 NVLink를 사용하여 1.8TB/s로 양방향 처리가 가능합니다.

NVIDIA Mission Control로 모델 실행과 핵심 업무 자동화

NVIDIA Mission Control은 세계 최고 수준의 운영 역량을 소프트웨어 형태로 조직에 제공하여, 개발자 워크로드부터 인프라, 시설 관리까지 AI 팩토리 운영의 모든 측면을 제어하는 플랫폼입니다. 이를 통해 훈련과 추론의 즉각적인 민첩성을 높이는 동시에, 인프라 복원력을 위해 풀스택 인텔리전스를 제공합니다. NVIDIA Mission Control은 어떤 기업이든 하이퍼스케일 수준의 효율성으로 AI를 구동하여 AI 실험을 가속화할 수 있도록 합니다. 또한, AI 개발과 배포를 간소화하는 소프트웨어 제품군인 NVIDIA AI Enterprise가 NVIDIA DGX 시스템에 최적화되어 있으며, NVIDIA NIM™ 마이크로서비스를 사용하여 속도, 사용 편의성, 관리 용이성, 보안성과 함께 최적의 모델을 배포할 수 있습니다.



최첨단 AI 팩토리 소프트웨어 스택

BN I&C



Technical Specifications

	DGX GB200
FP4 AI	1,440 PFLOPS
FP8 AI	720 PFLOPS
FP16 AI	360 PFLOPS
GPU	72x NVIDIA Blackwell GPUs in Grace Blackwell Superchips
GPU Memory HBM3e	13.5TB
Total Fast Memory	30.2TB
Interconnect	72x OSFP single-port NVIDIA ConnectX®-7 VPI with 400Gb/s NVIDIA InfiniBand 36x dual-port NVIDIA BlueField®-3 VPI with 200Gb/s InfiniBand and Ethernet
NVIDIA NVLink Switch System	9x L1 NVIDIA NVLink Switches
Management Network	Host baseboard management controller (BMC) with RJ45
Software	NVIDIA AI Enterprise: Optimized AI software NVIDIA Mission Control: AI data center operations and orchestration with NVIDIA Run:ai technology NVIDIA DGX OS / Ubuntu / Red Hat Enterprise Linux / Rocky: Operating system
Enterprise Support	Three-year Enterprise Business-Standard Support for hardware and software

Ready to Get Started?

To learn more about NVIDIA DGX GB200, visit:
nvidia.com/dgx-gb200

© 2025 NVIDIA Corporation and affiliates. All rights reserved. NVIDIA, the NVIDIA logo, Mission Control, BlueField, ConnectX, DGX, DGX SuperPOD, and NVLink are trademarks and/or registered trademarks of NVIDIA Corporation and affiliates in the U.S. and other countries. Other company and product names may be trademarks of the respective owners with which they are associated. 3432352. MAR25

